Library Technology Reports / Data Processing

June 1, 1964

THE MEDLARS PROJECT

By Joseph Becker

This is the first of an informal series of articles which the Library Technology Project is making available to the Bulletin. *Mr. Becker is an engineer and a librarian. He has had wide experience with governmental and commercial agencies as a consultant on data processing equipment, including service to ALA on the Library 21 and Library/USA World's Fair exhibits in Seattle and New York. His* Information Storage and Retrieval, *written with R. M. Hayes, was published late in 1963.*

Future articles will describe one or more significant computer projects in libraries, and may include a round-up of new developments. Since the Library Technology Project maintains a clearinghouse of information in this area, librarians are invited to send news items on their proposed plans or programs for use of computers and other data processing equipment to David Hoffman of the LTP staff at ALA headquarters. Another ALA unit interested in developments in this area is the Interdivisional Committee on Documentation of which Maurice F. Tauber of the Columbia University library school is chairman.

There are two main classes of computers—analog and digital. *Webster's Third New International Dictionary* defines the analog computer as a "type of calculating machine that operates with numbers represented by directly measurable quantities (as voltages, resistances, or rotations)," and the digital computer as "a computer that operates with numbers expressed directly as digits in a decimal, binary, or other system."

It is extremely unlikely that the analog computer will ever find a home in the library. Analog computers are generally used to control industrial processes, missiles in flight, etc. In oil refineries, for example, they supply the necessary automation for the simultaneous adjustment of thousands of valves. These valves regulate the kind and quantity of raw material which must flow through the jungle gym of pipes in order ultimately to produce gasoline of uniform quality.

To fix the two types of computers clearly in mind, one can think of the slide rule as representing the analog computer and the Chinese abacus as representing the digital computer. The slide rule provides a graduated scale; the abacus is operated bead by bead. Since numbers may be coded to represent the letters of the alphabet in the digital computer, it holds the major promise for handling verbal data and therefore is apt to have the greater impact on library methodology.

Logic and programming

Although the inventors of the digital computer initially designed their machines to contend with burdensome forms of arithmetical data, they also made the machines capable of performing logical operations. This is the really unique feature of the digital computer because its "logic" gives it the ability to select one of a number of alternative procedures according to the outcome of some previous computation. It is capable of making comparisons between two numbers to decide if a "greater than," "less than," or "equal to" condition exists. Thus, by substituting coded numbers for letters of the alphabet, the digital computer can be programmed to make decisions based on comparisons of nonnumerical data as well. This ability to make comparisons gives the computer its greatest power in the field of information handling. With it, for example, we are able to identify and select verbal data; to alphabetize; to sort, merge, and list various groups of words.

Unfortunately, this extraordinary ability of digital computers is sometimes misunderstood and gives rise to the fanciful notion that they are endowed with near-human or even superhuman qualities. Computer people encourage this in their vocabulary: You "instruct" and "query" the machine in its own "language"; it "accepts," "differentiates," "searches its memory," "analyzes," and even "evaluates." This anthropomorphizing creates an aura of mystery about computers and seduces people into believing that the machines can "think."

It is true that man has evolved ways to make the digital computer perform alpha-

betical operations which closely resemble human activity. It is also true that the machine can digest more information than a man and process it faster and more accurately. But the parallel with human skills stops there. The notion that a computer can act of its own free will, even to the extent of originating ideas, is a false one. Every step in the operation of a computer is not "thought out" by the computer itself but is carefully directed by the human mind. A "thought" achieved by the computer is not a thought at all, but a complex interconnection of circuitry, predetermined by man to function in a specific way in processing a problem to solution. A programmer is taught the operations which a computer is capable of performing. With this knowledge he is able to prepare precise instructions, called a "computer program," that drive the computer through the operations required to complete a specific job, such as alphabetizing, selecting, or printing words in a given format on a sheet of paper.

The computer is the heart of any data-processing system because it has the ability to process alphabetical data, to follow logical rules laid down by the programmer, and to release the required results in a variety of ways. In addition, however, the computer is surrounded by peripheral equipment which helps to bring the alphabetical data in and carry it out. These pieces of equipment are called input/output devices. A magnetic tape unit, for example, is an input device that brings data into the computer. High-speed printers are output devices which permit us to read the results on paper.

One of the most interesting and successful computer applications in a library is MEDLARS, described here.

MEDLARS

MEDLARS is an acronym standing for Medical Literature Analysis and Retrieval System. Of all computer activities going on in the library world today, this one at the National Library of Medicine is coming closest to achieving full operational status. As such, it particularly merits the attention and interest of professional librarians.

Long before this particular system was originated, the literature in the field of medicine was being systematically analyzed and retrieved. The end products of this bibliographical effort have been two publications which are very familiar: *Index Medicus* and the annual *Cumulative Index Medicus*. As early as 1955, Frank B. Rogers and Seymour Taine at NLM began investigating the use of mechanical aids in the preparation of printed indexes. This pioneer work provided them with the background and experience which eventually led to the MEDLARS project.

Figure I shows a block diagram of the three major elements which have characterized the flow of work at NLM. What is new about MEDLARS is that a digital computer has been introduced to assist NLM in its technical processing, retrieval, and prepublishing operations.

The MEDLARS computer application is not intended to replace indexers but only to enlist the power of the computer in speeding up those clerical processing tasks normally involved after the professional has done his work. Of the three elements shown in Figure I, only the latter two involve the use of the computer. The front end of the processing pipeline, where the basic intellectual activity occurs, remains unchanged.

How data are processed

Briefly, MEDLARS operates as follows. Medical journals are received by an indexer who scans them for subject content. Subject headings taken from *Medical Subject Headings* (MESH) are assigned to each article and cataloging data are extracted according to prescribed rules for the *Index Medicus*. The indexer types the results of his work onto a "data sheet." The data sheet is the critical link between the indexer and the computer because it provides, according to rigorous format specifications, the sequence in which the indexer's work will ultimately be treated by the computer. MEDLARS makes provision for revising, proofreading, and correcting the data sheets before they are sent to a typist for conversion to machine language.

At this stage the typist prepares conventional catalog entries using a special typewriter which converts the letters of the alphabet into punched holes in a paper tape as a by-product of the typing effort. (This was also done under the previous manual system.) Because three punched holes represent coded numbers for letters of the alphabet, their form is at once compatible with and easily interpreted by the

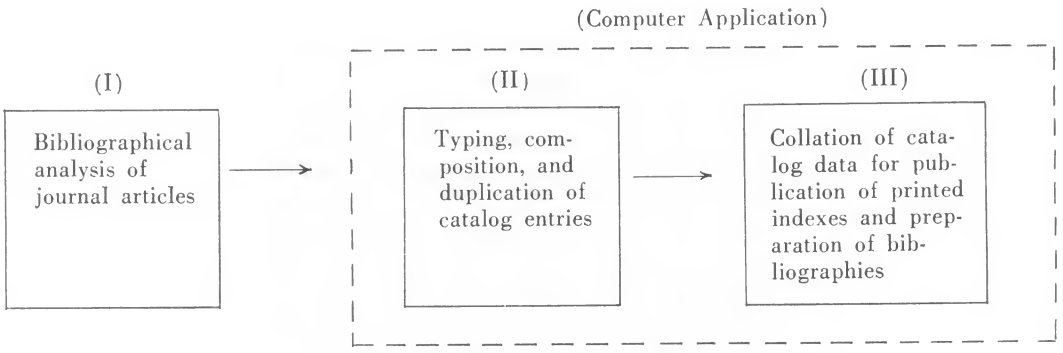

Fig. I—Routine Flow of Work

digital computer—hence the term "machine language." Production of machine language according to specifications is by far the most critical phase of the processing operation.

The final step in the conversion process occurs when the punched paper tapes are spliced together and run through a machine that transfers their intelligence to magnetic tape. The transfer procedure is quite simple—the holes in the punched paper tape are sensed and then faithfully duplicated as electrically charged spots on a spool of magnetic tape; MEDLARS input data are now ready for the computer.

The work the computer performs

A series of programs was written for MEDLARS that enables the computer to perform automatically many clerical operations previously done manually. For example, it can compare applied subject headings with a MESH authority file stored in memory to detect an invalid entry. It can generate full citations and assemble them as cross references in a master file. It can look for logical errors in the citations themselves and print out precisely where these occur. It can sort selected citations into alphabetic sequence. It can search its files to select information by any combination of elements which are contained in each citation, etc. Additional special computer programs also exist for processing requests for bibliographies and for generating lists of citations for eventual printing.

A unique machine called GRACE (Graphic Arts Composing Equipment) has been designed especially for MEDLARS; it is used peripherally to the main computer. This machine is capable of producing a positive film transparency ready for photo-offset printing.

GRACE is a one-of-a-kind machine for automatic electronic photo composition. Machines like it have existed for several years and are in common use for the automatic typesetting of newspaper advertisements. The advanced features of GRACE are its speed—30 times faster than predecessor machines—and its ability to utilize multiple type fonts to satisfy NLM's printing specifications.

In preparing the *Index Medicus* or the *Cumulative Index Medicus* for publication, the computer program first selects appropriate citations from its master file. Each citation contains the machine-language codes that are needed to make GRACE work. The program then organizes these citations into the order prescribed for publishing. All this results in output from the computer of a spool of magnetic tape containing the ultimate publication content and the instruction codes to operate GRACE. The output tape drives GRACE to produce automatically a positive film transparency. Citations on the film are properly justified, are grouped under their required subject headings, and appear in the various type fonts and sizes required to meet NLM's high-quality printing standards.

Many years of painstaking human effort on the part of the NLM professional library staff and their contractor, the General Electric Company, were required to conceive, design, and implement MEDLARS. Investigations first began in late 1960. This led to a preliminary design phase in late 1961 that took MEDLARS out of the concept stage and into the road-map stage essential for the detailed design phase that was to follow. The detailed design effort, begun in February 1962, lasted more than a year. During this time detailed performance specifications were prepared and

computer programs were written. The arrival of the computer did not occur until 1963 when a Honeywell 800 computer was installed and MEDLARS experienced its first shakedown. At this point NLM attempted to identify and exterminate as many of the conspicuous "bugs" as they could find. Several months later the manual system was replaced by the computer-driven system, except for GRACE. When GRACE is delivered in early 1964 total systems implementation will be completed.

MEDLARS objectives

Some librarians argue that MEDLARS constitutes an elaborate and expensive means of doing a job that traditionally has been done manually. On the other hand, others commend NLM for undertaking such pioneer work and view the effort as an important bench mark toward the achievement of more advanced goals. The following summarization of MEDLARS' objectives is to indicate whether the accomplishment of these objectives (which appears reasonably certain at this writing) will be worth the effort.

- Improve and enlarge *Index Medicus* and reduce the time required to prepare the monthly edition for printing from twenty-two to five working days.
- Produce other compilations similar to *Index Medicus* in form and content.
- Include citations derived from sources other than journal articles.
- Promptly (a maximum of two days) and efficiently service requests for special bibliographies on both a demand and a recurring basis, regularly searching at least five years of stored computer files.
- Increase the average depth of indexing per article by a factor of five, i.e., ten headings versus two.
- Nearly double the number of articles that may be handled annually—from the current 140,000 to 250,000 in 1969.
- Reduce the need for duplicative total-literature screening operations.
- Keep statistics and perform analyses of its own operations in order to provide the information needed to monitor and improve system effectiveness.
- Permit future expansion to incorporate new and as yet not completely defined objectives (e.g., communication of data from NLM to remote locations). •••

--Reprinted with permission from ALA BULLETIN, March, 1964.